AFFTC-PA-11030

Developing Statistically Defensible Propulsion System Test and Evaluation Techniques

David Kidman, Craig Stevens
Christopher Moulder, Dr. William Kitto,
Dr. James Brownlow, and Todd Remund

AIR FORCE FLIGHT TEST CENTER
EDWARDS AFB, CA

AFFTC

AIR FORCE FLIGHT TEST CENTER
EDWARDS AIR FORCE BASE, CALIFORNIA
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
15-04-2011	Technical Paper	N/A

4. TITLE AND SUBTITLE
DEVELOPING STATISTICALLY DEFENSIBLE PROPULSION SYSTEM TEST AND EVALUATION TECHNIQUES

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Kidman, David S., Propulsion Engineer
Moulder, Christopher J., Propulsion Engineer
Stevens, Craig A., Propulsion Engineer
Kitto, Dr. William, Statistician
Brownlow, Dr. James D., Statistician
Remund, Todd G., Statistician

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AND ADDRESS(ES)
773TS/ENFP
412 TW
307 E. Popson Ave.
Edwards AFB, CA 93524

8. PERFORMING ORGANIZATION REPORT NUMBER
AFFTC-PA-11030

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
773TS/ENFP
412 TW
307 E. Popson Ave.
Edwards AFB, CA 93524

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S) N/A

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release A: distribution is unlimited.

13. SUPPLEMENTARY NOTES
CA: Air Force Flight Test Center Edwards AFB CA CC: 012100

14. ABSTRACT

Acquisition of military hardware typically proceeds from design, development, production, and finally to operational use and support. Prior to the full-rate production, both developmental and operational test and evaluation (DT&E and OT&E respectively) must occur to ensure that the system meets military requirements. The United States Air Force (USAF) is continually looking for ways to improve its test and evaluation techniques. Since 1997, Air Combat Command (ACC) has been successfully using Design of Experiments (DOE) to construct and analyze operational test efforts. This paper highlights recent efforts to pursue statistically defensible test techniques to aid developmental test efforts.

Defensible testing is a statistical approach similar to DOE but emphasizes the need for better test planning by:

- insistence on understanding the system under test
- requiring clear and achievable test objectives
- ensuring system performance is measurable
- requiring that instrumentation accuracy and uncertainty propagation are well understood
- and requiring confidence, power, and performance thresholds

This paper highlights the Air Force Flight Test Center's (AFFTC) first steps to improve aircraft propulsion system test and evaluation (T&E) through the implementation of statistically defensible test techniques. Background on the AF acquisition process, the Air Force vision for defensible testing, and an aircraft propulsion T&E case study are presented.

15. SUBJECT TERMS
defensible testing (DOE Design of Experiments (T&E)Test and Evaluation confidence statistical analysis statistics
power sample size augmentor time-to-max linear model parallel lines

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
Unclassified					412 TENG/EN (Tech Pubs)
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*
Unclassified	Unclassified	Unclassified	None	7	661-277-8615

Proceedings of GT 2012
ASME Turbo Expo 2012: Power for Land, Sea, and Air
June 11-15 2012, Copenhagen, Denmark

GT2012-XXXX

DEVELOPING STATISTICALLY DEFENSIBLE PROPULSION SYSTEM TEST AND EVALUATION TECHNIQUES

David Kidman, Craig Stevens, and Christopher Moulder
US Air Force/773rd Test Squadron
Propulsion Integration Flight
Edwards AFB, CA
david.kidman@edwards.af.mil

Dr William Kitto, Dr James Brownlow, and Todd Remund
US Air Force Flight Test Center
Statistical Consultants Flight
Edwards AFB, CA
william.kitto@edwards.af.mil

ABSTRACT

Acquisition of military hardware typically proceeds from design, development, production, and finally to operational use and support. Prior to the full-rate production, both developmental and operational test and evaluation (DT&E and OT&E respectively) must occur to ensure that the system meets military requirements. The United States Air Force (USAF) is continually looking for ways to improve its test and evaluation techniques. Since 1997, Air Combat Command (ACC) has been successfully using Design of Experiments (DOE) to construct and analyze operational test efforts. This paper highlights recent efforts to pursue statistically defensible test techniques to aid developmental test efforts.

Defensible testing is a statistical approach similar to DOE but emphasizes the need for better test planning by:

- insistence on understanding the system under test
- requiring clear and achievable test objectives
- ensuring system performance is measurable
- requiring that instrumentation accuracy and uncertainty propagation are well understood
- and requiring confidence, power, and performance thresholds

This paper highlights the Air Force Flight Test Center's (AFFTC) first steps to improve aircraft propulsion system test and evaluation (T&E) through the implementation of statistically defensible test techniques. Background on the AF acquisition process, the Air Force vision for defensible testing, and an aircraft propulsion T&E case study are presented.

USAF ACQUISITION PROCESS

The U.S. Department of Defense (DoD) acquisition policy exists to manage the nation's investments in technologies, programs, and product support necessary to achieve the national security strategy and support the U.S. Armed Forces (Reference 1). The primary objective is to acquire quality products that satisfy users' needs with measurable improvements to mission capability and operational support in a timely manner and at a fair and reasonable price. The acquisition policy states, "Test and Evaluation shall be integrated throughout the defense acquisition process." The key here is to integrate T&E efforts across the project lifecycle using similar approaches. ACC has successfully promoted the use of DOE to construct and analyze operational testing; it is the intent of this paper to demonstrate applicability of this process to developmental testing.

The fundamental purpose of T&E efforts is to provide knowledge to manage the risks involved in developing, producing, operating, and sustaining systems and capabilities. T&E provides knowledge of system capabilities and limitations to the acquisition community for use in improving the system performance and the user community for optimizing system use and sustainment in operations. T&E enables the acquisition community to learn about limitations (technical or operational) of the system under development, so that they can be resolved prior to production and deployment to ensure systems are operationally mission capable (i.e. effective and suitable).

Figure 1 summarizes the acquisition lifecycle. Key T&E activities for each phase of acquisition are shown. These include early tester involvement and standup of integrated test teams (ITTs) prior to completion of the concept refinement phase. The T&E strategy is generally established prior to the technology development phase, and the test and evaluation master plan (TEMP) is established just prior to the System Development Stage (SDD) phase and is refined as necessary during project execution. As shown in Figure 1, integrated Government T&E (a combination of DT&E and OT&E) continues throughout the acquisition lifecycle.

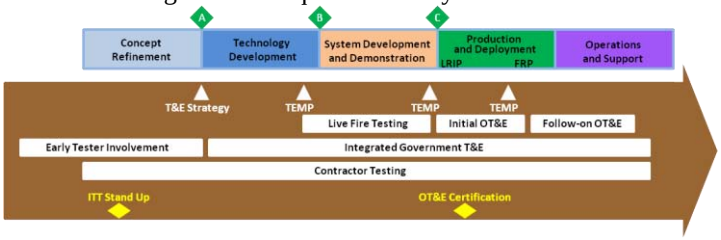
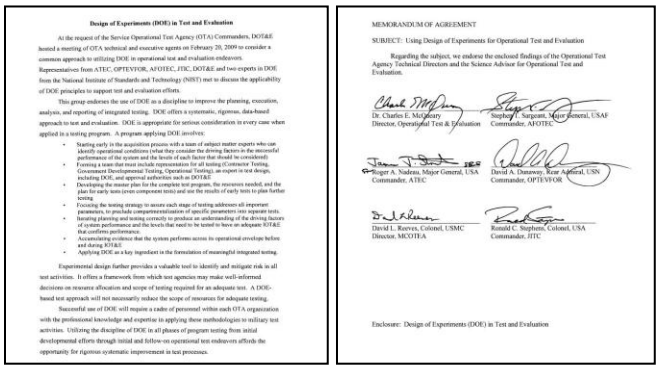


Figure 1. Integration of the Acquisition and T&E Processes

Recently, the DoD Director of OT&E signed a memorandum regarding the use of DOE as a discipline to improve test rigor. The memorandum shown in Figure 2 states, *"DOE provides the scientific and statistical methods needed to rigorously plan and execute tests and evaluate their results. DOE should allow DOT&E to make statements of the confidence levels we have in the test results. Whenever possible, our evaluations must include a rigorous assessment of the confidence level of the test, the power of the test and some measure of how well the test spans the operational envelope of the system"* (Reference 2).


Design of Experiments (DOE) in Test and Evaluation

At the request of the Service Operational Test Agency (OTA) Commanders, DOT&E hosted a meeting of OTA technical and executive agents on February 20, 2009 to consider a common approach to utilizing DOE in operational test and evaluation endeavors. Representatives from ATEC, OPTEVFOR, AFOTEC, JTIC, DOT&E and two experts in DOE from the National Institute of Standards and Technology (NIST) met to discuss the applicability of DOE principles to support test and evaluation efforts.

This group endorses the use of DOE as a discipline to improve the planning, execution, analysis, and reporting of integrated testing. DOE offers a systematic, rigorous, data-based approach to test and evaluation. DOE is appropriate for serious consideration in every case when applied in a testing program. A program applying DOE involves:

- Starting early in the acquisition process with a team of subject matter experts who can identify operational conditions (what they consider the driving factors in the mission/of performance of the system and the levels of each factor that should be considered)
- Forming a team that must include representation for all testing (Contractor Testing, Government Developmental Testing, Operational Testing) and an expert in test design, including DOE, and approval authorities such as DOT&E
- Developing the master plan for the combined test program, the resources needed, and the plan for early tests (even component tests) and not the results of early tests to plan further testing
- Executing the testing strategy to assure each stage of testing addresses all important parameters, to preclude compartmentalization of specific parameters into separate tests
- Iterating planning and testing correctly to produce an understanding of the driving factors of system performance and the levels that need to be tested to have an adequate IOT&E that confirms performance.
- Accumulating evidence that the system performs across its operational envelope before and during IOT&E
- Applying DOE as a key ingredient in the formulation of meaningful integrated testing.

Experimental design further provides a valuable tool to identify and mitigate risk in all test activities. It offers a framework from which test agencies may make well-informed decisions on resource allocation and scope of testing required for an adequate test. A DOE-based test approach will not necessarily reduce the scope of resources for adequate testing.

Successful use of DOE will require a cadre of personnel within each OTA organization with the professional knowledge and expertise in applying these methodologies to military test activities. Utilizing the discipline of DOE in all phases of program testing from initial developmental efforts through initial and follow-on operational test endeavors affords the opportunity for rigorous systematic improvement in test processes.

MEMORANDUM OF AGREEMENT

SUBJECT: Using Design of Experiments for Operational Test and Evaluation

Regarding the subject, we endorse the enclosed findings of the Operational Test Agency Technical Directors and the Science Advisor for Operational Test and Evaluation.

Dr. Charles E. McQueary, Director, Operational Test & Evaluation
Stephen J. Sargeant, Major General, USAF, Commander, AFOTEC
Roger A. Nadeau, Major General, USA, Commander, ATEC
David A. Dunaway, Rear Admiral, USN, Commander, OPTEVFOR
David L. Reeves, Colonel, USMC, Director, MCOTEA
Ronald C. Stephens, Colonel, USA, Commander, JITC

Enclosure: Design of Experiments (DOE) in Test and Evaluation

Figure 2. DOT&E Memo Requiring DOE approach in Operational Testing

As a result of this guidance, DOT&E test strategies are required to address DOE concepts. This guidance is currently being considered for the DoD Defense Acquisition Guidebook (DAG), which is designed to complement the DoD 5000 directive (Reference 3). The DAG acts as a guide for both DT&E and OT&E practitioners and is expected to state, *"A robust and properly integrated T&E program is an affordable and efficient program. Design of Experiments is a proven approach to apply in planning, executing and analyzing tests. Design of Experiments supports the development of effective and efficient test programs and will provide Program Managers (PMs) with measures of 'goodness' to assess tests and test programs."* The DAG guidance on DOE will also provide the PMs with the basic understanding needed to work with DOE experts who can apply modern scientific and statistical methods to test planning and execution. The intended outcome is an integrated T&E program in which contractor and government testers work from a common statistically relevant test plan.

DEFENSIBLE TEST APPROACH

Statistically defensive testing is an approach that emphasizes the need for test planning that includes clear and achievable test objectives, ensures system performance is measurable, insists on understanding the system under test, and requires that instrumentation accuracy and uncertainty propagation are well understood. This paper highlights the AFFTC first steps to improve aircraft propulsion system T&E through the implementation of defensible test techniques via a statistical analysis approach. In order for a statistical test approach to be successful, three key elements must be in place. The first and most important is to have clear and achievable test objectives with quantitative, mission-oriented metrics. The second is to describe how well the system's operational envelope is covered by testing, and the third is to calculate the confidence and power of the test.

MEASURABLE TEST OBJECTIVES

The first element in effective test planning is to identify the test objective. In this regard, determining the question to be answered helps guide the selection of the appropriate performance metrics, the test approach, and the applicable methodology. Performance metrics need to be observable, measurable, and testable. In many cases, they are limited in one or more of these areas. As an example, our case study examines thrust response differences caused by a revised digital engine control installed in a modern afterburning fighter type aircraft. The original test requirement stated that thrust response would be "considered satisfactory if time-to-max thrust was comparable to the previous engine control logic version." However, the requirement did not specify how thrust response was to be determined (small enough for pilot not to notice or large enough to have an operational impact) or where in the flight envelope thrust response should be compared (takeoff flight regime, cruise conditions, or across the entire aircraft flight envelope). As a result, for the original testing, thrust response for similar throttle transients (e.g. IDLE-MAX) was compared across the flight envelope.

In our case study, examples of measurable responses included: time for augmentor light-off and time for engine thrust response. Other measurables might have included engine stability, maximum exhaust gas temperature, or operational impact. However, these responses were not chosen because they did not directly influence thrust response. Additionally, operational impact frequently does not have a specific threshold (e.g. 10-percent degraded thrust response) that will impact flight operation.

As a side note, in cases where quantitative metrics are not available or practical, qualitative metrics such as surveys can be used. Surveys typically contain quantitative

information that can be summarized with statistical measures and qualitative information in written format. There are many ways to gather quantitative data from these surveys (e.g. Cooper-Harper ratings of thrust response impacts to aircraft handling qualities). If executed correctly, surveys can be a valuable resource; however, care must be taken in designing surveys to ensure that the maximum amount of information is obtained.

ENVELOPE COVERAGE

The second element in defensible test planning is to ensure that the planned test adequately covers the operational envelope. Only operationally realistic flight conditions should be considered, so judgment is critical. It is also important to document limitations that prevent testing specific combinations of factors that drive performance (e.g. altitude, airspeed, or engine health). The number of factors can influence system performance can vary widely from one system to another and all significant factors should be included. For each factor, levels of operation must also be examined. For example, if airspeed is a factor that drives thrust response time, important speed regimes should be identified (e.g., low speed, transonic and supersonic). Test events that address interactions of factors (e.g. airspeed and altitude) should also be considered.

Claims with respect to system performance in areas where the system was not tested should not be made. In our case study, no indication of operational envelope coverage was specified at the time of original test planning. The original case study test matrix (Figure 6) was developed to evaluate thrust response and engine stability, focusing on the most challenging areas (e.g. upper left hand corner of the aircraft flight envelope) and logic implementation area. Tests were also planned based on knowledge of system operation gained from past testing. In retrospect, adequate envelope coverage would have had test points that more equally spanned the entire range of engine inlet pressure (PT2) and temperature (TT2). In our case study, testing across the entire flight envelope was not required, as engine operation was not expected to change at supersonic flight conditions. As a result, only a spot check of system performance was required at supersonic flight conditions

CONFIDENCE AND POWER

The third element in defensible test planning is to consider test confidence and power. In the planning stages, the acceptable risk, power, and sample size must be clearly documented. Power is the probability that a test will capture a difference between two datasets if it exists. While confidence is the probability that the prediction is correct. The confidence level needed for the test should be determined during pre-test planning and all results should be displayed with confidence values. Typical values are 95-percent or greater depending on the level of risk the test team is willing to accept.

Power is influenced by sample size (amount of testing), noise (standard deviation), and the performance threshold testing should capture. Generally, the greater the sample size the greater the power. However, this assumes the distribution of the tests points adequately captures the behavior of the system. Without prior knowledge, a common test execution strategy is simply an even spacing of the points along the variable of interest (e.g. PT2). Also during test planning, the tester typically does not know the signal-to-noise ratio of the test data. The signal is the detectable difference in performance (e.g., time-to-max). Noise is a function of system repeatability and measurement error and is measured in terms of the data's standard deviation. Not surprisingly, it is easier to detect signals that are much larger than the noise. By assuming notional signal to noise (S:N) ratios, the analyst can compare power across different test approaches without making assumptions about the data. The power numbers shown in the notional chart in Figure 3 provide an example of this.

It should be noted that a test with high power does not equate to a high probability of capturing aberrant behavior; it only speaks to probabilities related to well behaved systems. If there is a chance that a system will perform differently in a small region, the goal of the test must be reanalyzed, and a companion test considered. Power levels above 80-percent are typically considered sufficient when failure is not life threatening or expected to cause significant financial burden.

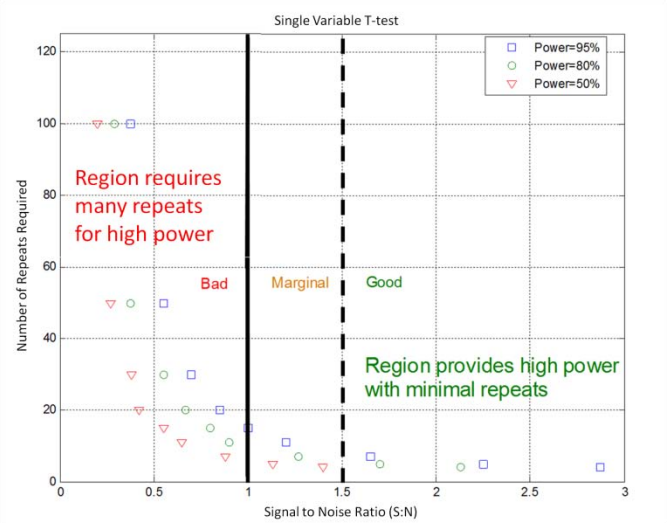


Figure 3. Notional chart showing impact of signal to noise ratio on testing required

CASE STUDY: ENGINE CONTROL UPGRADE FOR AFTERBURNING FIGHTER TYPE AIRCRAFT

Test Item Description

As part of the aircraft engine Component Improvement Program (CIP), the USAF Aeronautical Systems Center requested that the AFFTC evaluate a military fighter aircraft with a low-bypass turbofan engine and a revised digital engine control. The engine control was revised to increase stall margin in the heart of the envelope by increasing compressor variable vane camber (Figure 6). These changes were not expected to significantly degrade engine thrust response.

Source: http://www.aerospaceweb.org

Figure 4. Various military fighter aircraft

Source: http://www.aircraftenginedesign.com/

Figure 5. Various military low-bypass turbofan engines

The original test objective was to "Compare afterburner throttle transient capability, specifically compare time-to-max, to the legacy engine control logic." The objective was fairly vague and failed to specify how time-to-max should be determined or what would be considered successful.

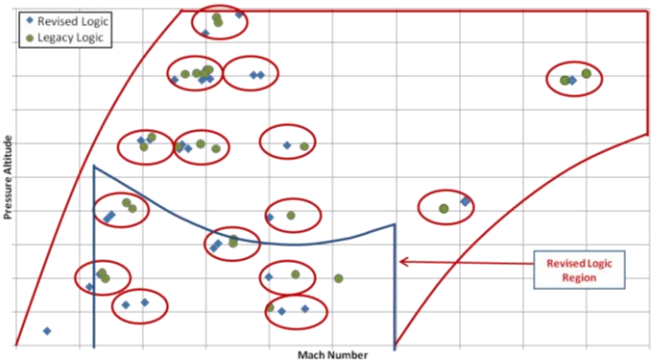


Figure 6. Flight test matrix to evaluate effects of the revised engine control

When the original flight test occurred, the analysis assessed revised logic effects using a variety of approaches. The first method included qualitative comparison of key engine parameters with both the legacy and the revised logic overlaid (see Figure 7). Since only a limited number of flight points were analyzed in this manner, thrust response changes across the flight envelope were difficult to observe.

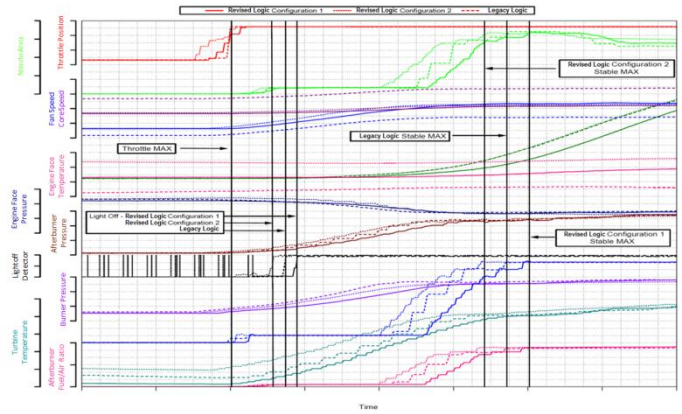


Figure 7. Historical analysis approach showing time history comparison plot of IDLE-MAX throttle transient

The second method used to determine revised logic effects was to compare average time-to-max for similar throttle transients (e.g. IDLE-MAX) across all flight conditions tested, as shown in Figure 8. This approach ignored the possibility that the revised control logic dataset had the same time-to-max values as the legacy, but at different flight conditions. In retrospect, adding 99% confidence bounds to this data did not make the results any clearer. Furthermore, given that confidence intervals are intended to speak to the entire population, it was probably incorrect to apply such bounds to data gathered from only two planes. The original evaluation concluded that there was no significant difference in IDLE-MAX thrust response between the revised and legacy logics.

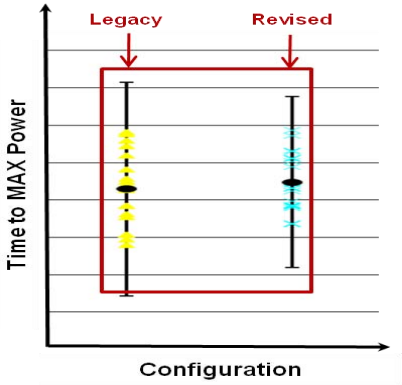


Figure 8. IDLE-MAX thrust response comparison with revised and legacy engine control logic

New Defensible Approach

The intent of implementing a defensible test approach is to apply a more rigorous statistical analysis process. The goal being to ensure that test objectives were achievable with measureable mission-oriented metrics, that testing adequately covered the system's intended envelope, and that confidence and power of final test results are known.

Measurable Test Objectives

As previously noted, the original test objectives in the case study did not precisely state how operability or time-to-max thrust would be determined or what was considered successful. It would have been better to re-state the test objectives in a defensible manner, such as *"Determine with statistical confidence that the revised engine control logic IDLE-MAX thrust response has not degraded in the revised logic implementation regime as compared to legacy engine control logic."* A similar objective could be written for flight regimes outside the logic affected flight regimes or focused on other types of throttle transients (e.g. IDLE-MIL or MIL-MAX).

Envelope Coverage and Power

The original test matrix was based on the goal of being able to compare revised results to legacy test results and focused on the most challenging areas for engine stability and regions where the engine control logic had been revised. In order to determine envelope coverage and power for the new defensible test objectives, the test matrix was divided into regions both inside and outside the revised control logic region. To analyze the power of the test for the revised defensible objectives an Analysis of Covariance (ANCOVA) was performed. The ANCOVA contained the output variable time-to-max thrust and two input variables, PT2 and the categorical variable (revised engine control logic). After removing the variance due to PT2, the ANCOVA tested whether the revised engine control logic affected the time-to-max thrust output variable.

The calculated power of the test executed for the region inside the revised control logic region was 0.97. This included sample sizes of 10 test points with the revised logic and 8 test points with the legacy logic. The power of the test performed outside of the revised control logic region was 0.98, which included 18 test points with the revised logic and 16 with the legacy logic. Even though testing outside the revised control logic region had significantly more data, the increased variation of results in this region resulted in little change in overall power. In both cases, the test executed exhibited adequate power for the difference observed, since test power was greater than 80-percent. In retrospect, it would have been possible to perform fewer test points and retain adequate test power. Assuming adequate envelope coverage, the number of points inside the revised control logic region could have been reduced from 18 to 10 (5 with the revised logic and 5 with the legacy, a 44-percent reduction), and the number of test points outside the revised control logic region could have been reduced from 34 to 20 (10 with the revised logic and 10 with the legacy, a 41-percent reduction). In many cases, a 10-percent difference is considered operationally significant. If the test was re-designed to capture a 10-percent difference, significantly more testing would have been required (46 test points inside the revised control logic region and 36 points outside the revised control logic region) to attain 80-percent power.

Confidence

While a confidence interval is often useful when describing how small or large a difference is, it is not always practical. In this case study, it was omitted for two reasons. First, because of a necessary logarithmic transformation on the data, the resulting confidence interval had reduced meaning. In other words, the log transformation added so many exceptions and caveats to the interval, that the ability of the common user to apply it to an operational scenario was greatly reduced. Second, because of the limited number of aircraft that data was gathered from, it was difficult to show that a confidence interval would have any interpretive ability beyond the aircraft tested.

However, other measures of confidence were available. In this study, our test confidence was related to the p-value, which statistically showed that a difference in time-to-max thrust existed. P-values lower than 0.05 (Figure 9) are indicative of a difference being present, but in contrast to a confidence interval, it does not tell you how great the difference is (Reference 4).

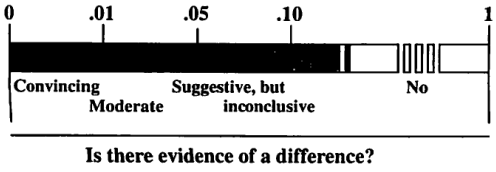


Figure 9. p-value

The statistical analysis approach needed to recognize that time-to-max thrust was a function of more than one input variable, and that one-dimensional analyses were not adequate to compare data at multiple flight conditions. To make such a comparison, it was necessary to define one or more variables that most influenced the engine operating state. Furthermore, it was desired that such a parameter be operationally relevant to allow easy application when determining impacts.

Two methods were used to determine which variables most impacted the engine time-to-max thrust evaluation criteria. One was an automated regression approach, which evaluated all available influence parameters (e.g. altitude, Mach, inlet pressure/temperature) and determined if direct or indirect mathematical relationships could be established. Another simpler technique was drawing on past experience and engineering judgment to establish how the data should trend. In typical turbine engine operation, engine inlet pressure is almost always a primary input to the control system and as a result, makes an ideal first guess when looking for inputs which most influence engine operation. Figure 10 shows time for the engine to achieve max power as a function of engine inlet pressure.

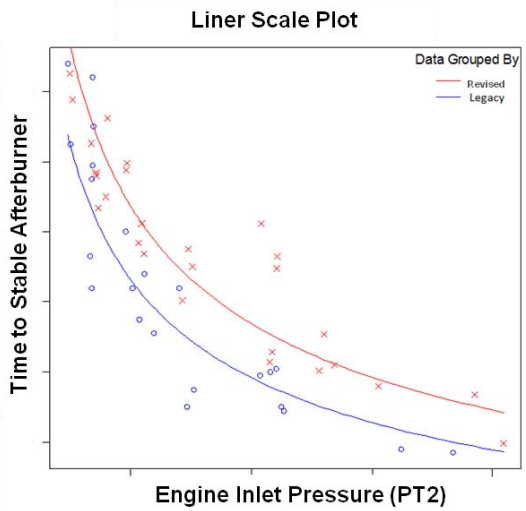


Figure 10. IDLE-MAX thrust response comparison with revised and legacy engine control logic

In order to keep the modeling effort simple and operationally relevant, a model of time-to-max thrust was developed as a function of PT2 and engine logic. The simplest approach was to model with a linear relationship using two coefficients:

$$y = mx + b \qquad \text{or} \qquad \textbf{Eq. 1}$$
$$Time\ to\ max = \beta_0 + \beta_1(PT2)$$

In Equation 1, $\beta_0$was the y-intercept and $\beta_1$the slope (Figure 11). This model was easily modified to include a third factor, $\beta_2 * Logic$. In this case, $\beta_2 = 0$ represented the revised logic and $\beta_2 = 1$ the legacy logic. Therefore, if β_0 and β_1 were held the same for both sets of data, the vertical shift between lines would capture the increase in thrust response time caused by the *Logic* parameter.

To generate a linear relationship, the data needed to be logarithmically transformed (Figure 11). These transformations are generally used to condition data that does not appear well behaved. Several such transformations exist, but each must be applied according to different criteria. A log transform is typically applied when data does not appear normally distributed, when the variances of the datasets being compared are not similar, or when the ratio of the largest to smallest measurement in the group is greater than 10. In this instance, a log transform was chosen only because of the exponential behavior of the parameters. Taking the log transform of both axes resulted in a linear relationship between PT2 and time-to-max, making the conditioned data well suited for a parallel-lines model statistical analysis technique.

$$\ln(Time_{ABStable}) = \beta_0 + \beta_1 ln(PT2) + \beta_2 Logic \qquad \textbf{Eq. 2}$$
$$Time_{ABStable} = e^{\beta_0} + (PT2)^{\beta_1} + e^{\beta_2 Logic}$$

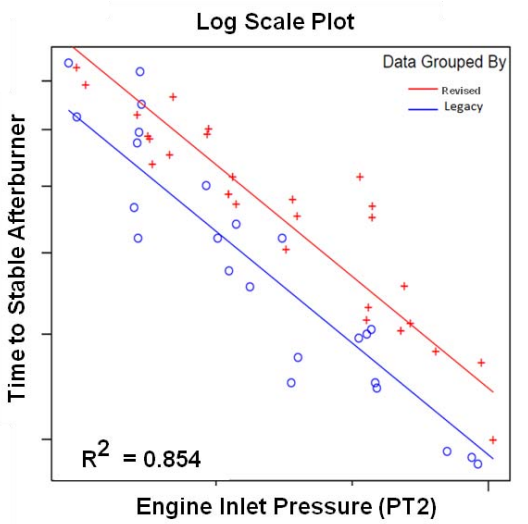


Figure 11. IDLE-MAX thrust response comparison with revised and legacy engine control logic (Log Scale)

A few drawbacks exist for logarithmic models. First, we can no longer talk about mean values, only the medians. This arises from the fact that the exponent of the mean of a set of log values is not equal to the mean of the actual values. However, because the rank order of individual values are the same in log space, comparisons can be made about the data median. Second, the *Logic* parameter, $\beta_2 * Logic$, has a multiplicative instead of an additive effect, which increases the amount of worked required to determine the absolute time difference between logics.

$$Time_{ABStable} = e^{\beta_0}(PT2)^{\beta_1} e^{\beta_2 * 0} = e^{\beta_0}(PT2)^{\beta_1} \qquad \textbf{Eq. 3}$$
$$Revised\ Logic$$
$$Time_{ABStable} = e^{\beta_0}(PT2)^{\beta_1} e^{\beta_2 * 1} = e^{\beta_0}(PT2)^{\beta_1} e^{\beta_2}$$
$$Legacy\ Logic$$

The value at PT2 for the legacy logic was scaled by the factor e^{β_2}. Therefore, the difference between logics was represented as a percentage difference. It had to be multiplied by an existing time value at a given PT2 to determine the actual time difference.

The parallel lines model shown in Equation 2 requires that a number of conditions be met. First, the variables of interest had to display a linear relationship. Second, the interaction between the independent variable (PT2) and the logic variable had to be minimal, thus ensuring that both lines had the same slope. Third, both sets of data had to independently show a strong correlation to the model. In this case study, a high R^2 (0.854) was used as evidence of a strong correlation. R^2 is a measure between 0 and 1 that describes the proportion of variance explained by a model.

As a result of this analysis, for data inside the revised control logic region (Table 1), the β_2 coefficient was -0.26180 and had a p-value of .000514. Converting the -0.2618 logarithmic units into the linear domain shows that the revised logic has a median increase of 30-percent more time to reach max thrust. Furthermore, the p-value of .000514 was very small, indicating high confidence that there was difference between datasets. For data outside the revised control logic

region, the median time-to-max thrust was determined to be 15-percent greater for the revised logic.

Outside the revised control logic region, thrust response was not expected to change. The difference may have been caused by engine to engine variation, aircraft installation effects, variation in flight conditions, or the engine control software itself. In our case study, the legacy data used as a baseline was just previously available data. If baseline testing had used the same engine and aircraft, then much of this uncertainty would not exist.

Table 1. Results of statistical analysis for test data residing inside the revised control logic region

	Estimate	Std. Error	t value	p-value
B_0	2.85750	0.23050	12.397	2.77e-09
B_1	-0.54009	0.09095	-5.938	2.72e-05
B_2	-0.26180	0.05946	-4.403	0.000514

Table 2. Results of statistical analysis for test data residing outside the revised control logic region

	Estimate	Std. Error	t value	p-value
B_0	2.54658	0.07393	34.447	< 2e-16
B_1	-0.45034	0.04321	10.422	< 2e-16
B_2	-0.14239	0.03883	-3.667	0.000946

CHALLENGES

A few challenges still exist that may influence the successful implementation of statistically defensible test techniques. These include: confounding variables, randomization, insufficient testing, and lack statistical knowledge among test engineers. It should be noted that depending on how the testing was performed, isolating the degraded thrust response caused by the revised engine control logic may be difficult. Confounding variables might include engine-to-engine variation (degradation from normal usage or manufacturing tolerances), variances in test conditions, or aircraft installation differences. Unless effects from these variations are known, care should be taken to ensure baseline results for comparisons are taken with the same engine and aircraft, ensuring variations between test data sets are minimized. Also, current propulsion test methodology includes the practice of testing one engine and making fleet decisions, which violates the scope of inference of the test. It is not statistically defensible to apply a finding to the fleet if a large enough subset was not tested. Another challenge is that classical DOE requires randomization during test execution to eliminate any impact of pilot learned response and control system hysteresis. However, efforts to minimize test costs and maximize test safety generally require testing to be performed in a methodical buildup fashion. Care must be taken, or the lack of randomization may affect the test outcome. Test budgets also limit the amount of testing possible, impacting statistical relevance and the ability to determine if a difference in engine or aircraft operation exists. Finally, engineers who plan test efforts are typically not statisticians. In order to minimize the effect of this last item, the AFFTC has recently initiated a Statistics Office whose goal is to provide statistical consultation on all matters concerning defensible test techniques and analysis.

SUMMARY

In summary, the Air Force is moving towards a policy requiring the use of statistically defensible test techniques. In applying a more rigorous approach to testing, it is important to understand that there is no “one size fits all” solution. Specific defensible methodologies depend on the system under test and the questions being asked. The primary benefit of the defensible approach is increased test rigor to ensure that a defensible test and evaluation approach is performed. This forces those involved in the T&E process (e.g. program managers, test engineers, and original equipment manufacturers) to finalize the true test strategy early. This strategy includes setting the test objectives, determining the inputs and interactions that may influence test outcome, and the measurables that will indicate success. Implementing defensible approaches highlights the opportunities that exist to optimize testing.

In summary, the statistical analysis was able to show with high confidence that the median revised engine control logic thrust response was slower. It was also noted that depending on how the test was performed, isolating the degraded thrust response to the revised engine control logic could be difficult due to confounding variables like engine-to-engine variation, variances in test conditions, or aircraft installation differences. As a result, care must be used to ensure comparable results are available. Furthermore, it was seen that the power of the test originally performed was excessively high (approximately 0.97). Analysis showed it would have been possible to perform approximately 40-percent less testing and still achieve acceptable power for determining confidence in test results.

REFERENCES

1. United States. Defense Acquisition University. DoD 5000 Series. 2008. 1 Jan 2011. <https://acc.dau.mil/CommunityBrowser.aspx?id=18532>.
2. McQueary, Charles E. Memo to DOT&E. 1 May 2009. <https://acc.dau.mil/CommunityBrowser.aspx?id=312213&lang=en-US6>.
3. United States. Defense Acquisition University. Defense Acquisition Guidebook. 2011. 1 Jan 2011. <https://dag.dau.mil/Pages/Default.aspx>.
4. Ramsey, Fred L. and Daniel W. Schafer. The Statistical Sleuth. 2nd ed. Belmont, CA: Brooks/Cole, 2002. 47.